A Thermal Infrared Video Benchmark for Visual Analysis

Zheng Wu
The Mathworks Inc.
Natick, USA

Nathan Fuller
Department of Biology
Boston University
Boston, USA

Diane Theriault, Margrit Betke
Department of Computer Science
Boston University
Boston, USA

***Abstract*—We hereby publish a new thermal infrared video benchmark, called TIV, for various visual analysis tasks, which include single object tracking in clutter, multi-object tracking in single or multiple views, analyzing motion patterns of large groups, and censusing wild animals in flight. Our data describe real world scenarios, such as bats emerging from their caves in large numbers, a crowded street view during a marathon competition, and students walking through an atrium during class break. We also introduce baseline methods and evaluation protocols for these tasks. Our TIV benchmark enriches and diversifies video data sets available to the research community with thermal infrared footage, which poses new and challenging video analysis problems. We hope the TIV benchmark will help the community to better understand these interesting problems, generate new ideas, and value it as a testbed to compare solutions.**

***Keywords*-thermal infrared benchmark; object detection; visual tracking;**

I. Introduction

The fast growth in computer vision research in the last decade has mostly been associated with visible-light sensors. Non-visible spectrum sensors have not been used as widely because, initially, low cost cameras had poor spatial resolution and a narrow dynamic range, and cameras with better image quality were prohibitively expensive for many researchers. Sensor technology has now advanced to a point that non-visible range sensors have regained researchers' attention in both academia and industry. With our work, we intend to answer the community's need for a comprehensive benchmark for a now popular non-visible range sensor, the thermal infrared camera. This passive sensor captures the infrared radiation emitted from the scene and its objects. Thermal imaging was originally developed for industrial and military use, for example, surveillance and night vision tasks. Recent studies have gone beyond the traditional tasks and applied thermal imaging to monitoring of wild animals, non-invasive food inspection, and heat loss detection [1], [2]. Our goal here is to provide the research community with a diverse set of video sequences that addresses various common computer vision problems. The proposed benchmark comes with a large number of high quality annotations to facilitate quantitative evaluations and comparisons of detection and tracking algorithms.

A few of thermal infrared dataset have been published in the past, e.g., the OTCBVS Benchmark [1], the LITIV Thermal-Visible Registration Dataset [3], the AIC Thermal-Visible Night-time Dataset [4], and the ASL Thermal Infrared Dataset [5] (Table I). Typically these datasets focus on specific biometric applications or involve thermal-visible multimodal systems and imply a close-up view of the objects in the scene. For general tasks, such as object detection and tracking, the usefulness of these datasets as benchmarks is limited due to their low image resolution, short video duration, and most importantly, lack of complexity of visual events in realistic, challenging environments. In contrast, our new thermal infrared video (TIV) dataset was collected by high-resolution high-speed cameras (FLIR SC8000, FLIR Systems, Inc., Wilsonville, OR), with a series of carefully designed recording protocols and preprocessing steps. The TIV benchmark covers five common computer vision tasks:

- Tracking a single object through clutter,
- Tracking multiple objects from a single view,
- Tracking multiple objects from multiple views,
- Visual counting,
- Group motion estimation.

In addition, background subtraction and object detection, are generally required as part of the solution. The categories of objects of interest, included in TIV, are pedestrians, marathon runners, bicycles, vehicles, and flying animals at various resolutions (see Fig. 1 for some snapshots). So far, TIV consists of 63,782 frames, recording thousands of objects; active updates are in progress. To the best of our knowledge, this is the largest thermal infrared video dataset available to the public.

II. TIV Dataset Description

Our TIV dataset consists of seven different scenes, two of them indoor scenes. Most of the data were recorded with FLIR SC8000 cameras (FLIR Systems, Inc., Wilsonville, OR), except sequences *Davis08-sparse*, *Davis08-dense*, *Davis08-counting*, which were previously published [8]. The full resolution is 1024×1024, but we use cropped images for some sequences in order to focus on regions of interest. Each pixel is described by 16 bits and has a value typically ranging

[1] http://www.vcipl.okstate.edu/otcbvs/bench/

Table I
SUMMARY OF THE PROPERTIES OF THE THERMAL INFRARED VIDEO TIV BENCHMARK

Data	Resolution	#Seq.	#Frames	Category	Density	Views
OTCBVS						
OSU Pedestrian [6]	360×240	10	284	Pedestrian	Sparse	1 (thermal)
OSU Color-Thermal [7]	320×240	6	17,089	Pedestrian	Sparse	1 (thermal) + 1 (visible)
IRIS Face	320×240	N/A	8,456	Face	N/A	1 (thermal) + 1 (visible)
Terravic Face	320×240	20	23,335	Face	N/A	1 (thermal)
Terravic Motion	320×240	18	25,355	Pedestrian	Sparse	1 (thermal)
Terravic Weapon	320×240	5	1,900	Divers, Plane Weapon	N/A	1 (thermal)
LITIV [3]	320×240	9	6,236	Pedestrian	Sparse	1 (thermal) + 1 (visible)
ASL-TID [5]	324×240	9	4,381	Pedestrian, Cat, Horse	Sparse	1 (thermal)
Our TIV	up to 1024×1024	16	63,782	Pedestrian, Runner Car, Bicycle Motorcycle, Bat	Sparse Medium Dense	up to 3 (thermal)

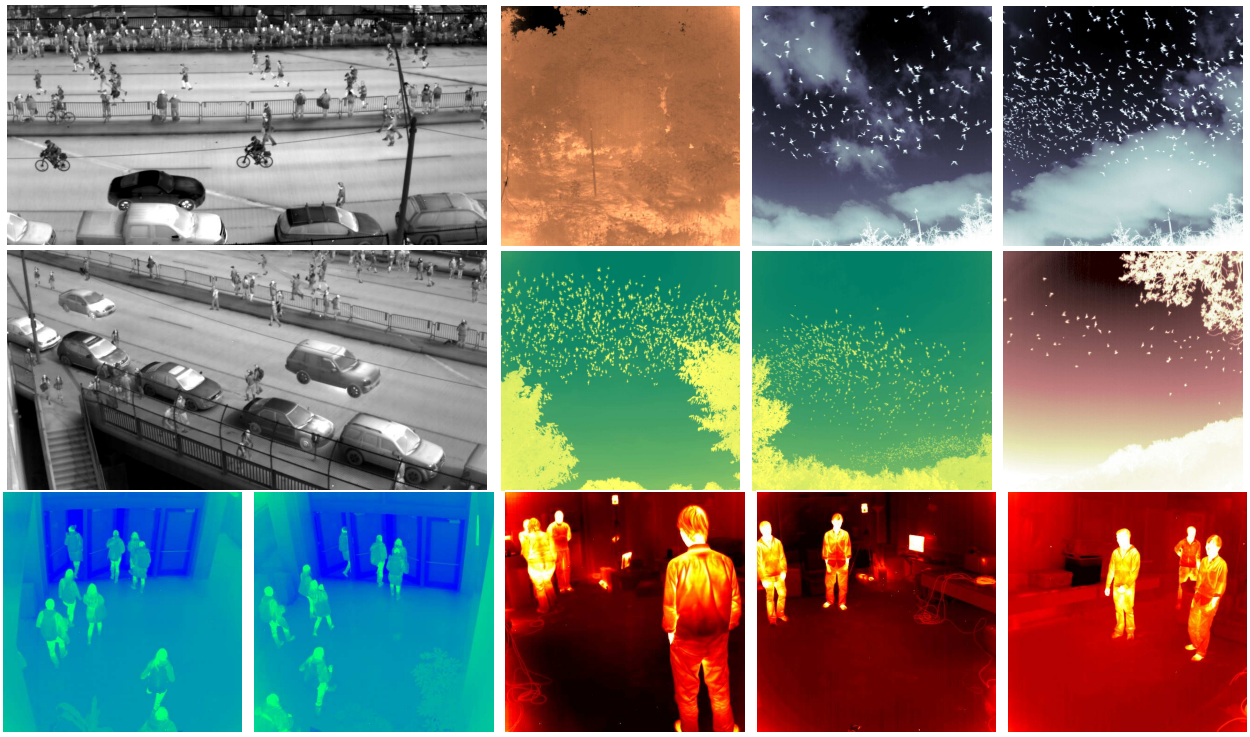


Figure 1. Snapshots of TIV dataset. Sequences captured from the same scene are grouped with the same false color.

between 3,000 to 7,000 units of uncalibrated temperature. The frame rate was set between 5 to 131 fps depending on the speed of the objects in the scene. The full list of sequences is given in Table. II.

Thermal cameras typically exhibit a fixed pattern of noise caused by the nonuniform response of the sensor across the pixel array. For the users' convenience, the benchmark includes both raw data and image data after we applied a "nonuniform two-point correction pre-process" [9], [10], in which two uniform sources of intensity ("cold" and "hot") were sequentially imaged. For each pixel, the difference between the measured intensity y_m and the corrected intensity y_c of the image is expressed as the linear approximation

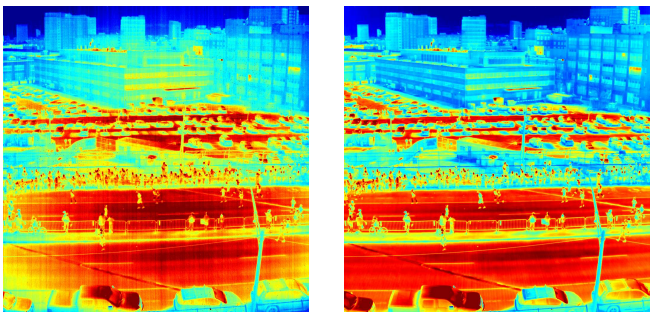
$$\Delta y = y_m - y_c = a \cdot y + b. \quad (1)$$

From the hot and cold measurements, the multiplicative gain a and additive offset b can be computed for each pixel. The output of the nonuniform correction is obtained by subtracting the approximated difference Δy from the original input y_m. An example image before and after nonuniform correction is given in Fig. 2.

Four out of seven scenes in TIV have multiview support.

Table II
SUMMARY OF THE VIDEO SEQUENCES IN OUR TIV BENCHMARK

Name	Resolution	#Fra.	Category	Density	Views
Atrium	512×512	7,964	People	medium	2
Velifer	1024×1024	3,000	Bat	Sparse	3
Bracken-counting	1024×1024	150	Bat	Dense	1
Bracken-flow	1024×1024	10,000	Bat	Dense	1
Davis08-sparse	640×512	3,300	Bat	Sparse	3
Davis08-dense	640×512	600	Bat	Dense	3
Davis08-counting	640×512	300	Bat	Dense	1
Davis13-medium	1024×1024	1,500	Bat	Medium	3
Frio10	1024×512	499	Bat	Dense	1
Frio11	1024×1024	299	Bat	Medium	1
Lab	512×512	26,760	People	Medium	3
Marathon-1	1024×512	1,000	Pedestr.	Medium	1
Marathon-2		2,999	Runner	Medium	1
Marathon-3		1,275	Bicycle	Medium	1
Marathon-4		1,282	Motorcy.	Medium	1
Marathon-5	1024×640	6,000	Car	Medium	1



(a) Raw Frame (b) Corrected Frame

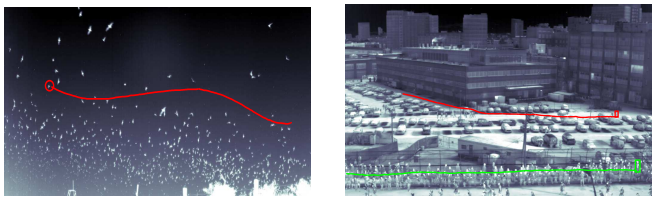
Figure 2. Sample images for nonuniform correction.

When multiple cameras were used, all cameras were synchronized with a signal generator that triggered the recording at the same time. TIV includes camera calibration information. For planar motion (*Atrium* and *Lab*), a homography-based ground plane was estimated [11]. For free motion in 3D space (*Velifer* and *Davis13-medium*), we applied the self-calibration procedure proposed by Theriault et al. [2].

In the following sections, we show the use of specific TIV sequences to address five different visual analysis tasks.

A. Tracking a Single Object through Clutter

Tracking a single object through clutter is one of the most active research topics in computer vision [12]. The task starts typically with a manual initialization to specify the object of interest. The object is then tracked throughout the sequence. The object may experience appearance changes, have interactions with distractors, or change its motion pattern. Most of the state-of-the-art algorithms focus on appearance modeling and the search strategy, and use



(a) *Frio10* (b) *Marathon-5*

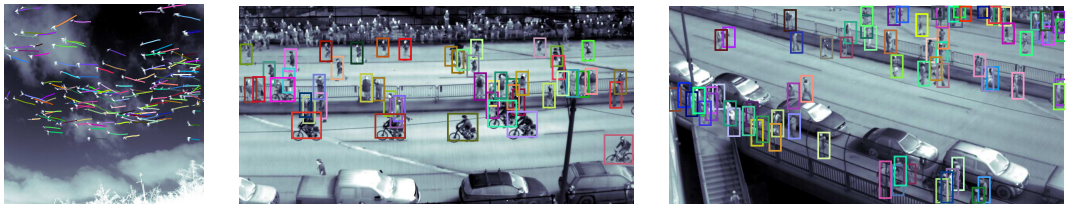
Figure 3. Sample frames for single object tracking.

machine learning tools. They typically cannot be directly applied to infrared videos, because there are other, unique challenges here. The thermal radiation helps the foreground object to stand out in the image, but very often the object also loses appearance details. Moreover, it is very difficult to distinguish multiple objects having the same thermal profile. To specifically highlight these two issues, we collected the sequences *Frio10* and *Marathon-5* (Fig. 3).

For the *Frio10* sequence, the task is to track 10 specified bats during the emergence of the colony. The density of the bats is high, while the resolution of each bat is small. There are frequent partial or complete occlusions, but the periodic motion pattern of each bat is relatively stable. For the *Marathon-5* sequence, the task is to track 10 specified pedestrians walking on busy sidewalks and between parked cars. The background is noisier in this case, and there are frequent occlusions as well. Given the small resolution of the objects in the image, we only annotated a single point for each object and smoothed the trajectory.

Baseline Method and Evaluation. To initialize a track, we used either the annotation from the first or the last frame. We call these"tracking forward" and "tracking backward" initializations. For the *Frio10* sequence, the baseline is a detection-based method that applies an object detector and filters the state of the object by a nearest neighbor search. The object detector requires background subtraction and localizes the objects by computing the pixels with local intensity maxima within each binary disconnected component. For *Marathon-5*, the baseline is an intensity-based method that uses normalized correlation to find the best match in the next frame. Both methods also apply a linear dynamic motion model to predict the position of the object when the detection fails or the correlation score is not sufficiently high.

To evaluate baseline performance, we computed the Euclidean distance between the tracked position and the ground truth in each frame. If the distance was smaller than a predefined hit/miss threshold, we claimed a good match was found. Throughout the experiments, we chose 5 pixels as the threshold. The key metric, "success rate," is defined as the total number of good matches divided by the total number of frames, with ideal value 1. We do not encourage the usage of the traditional metric "mean distance error" here for reasons:

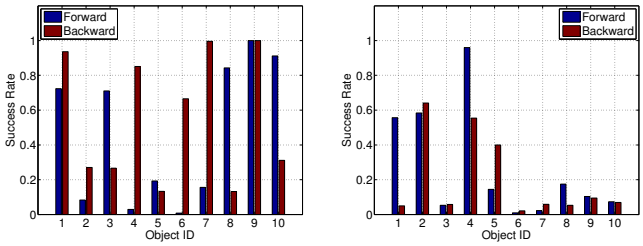


(a) *Frio11* (b) *Marathon-2* (c) *Marathon-4*

Figure 5. Sample frames for multi-object tracking from a single view.

1. The error can become arbitrarily large when the tracker drifts. 2. The resolution of the object in our experiment is small, so the tracked pixel location within the region of the object is not crucial. 3. The 5 pixel hit/miss threshold is sufficiently small to guarantee that the tracked position falls into an acceptable region of trust.

The results of the baseline methods on the two sequences are shown in Fig. 4. The average success rate is 51% for the *Frio10* sequence, and 23% for the *Marathon-5* sequence, which suggests there is much room for future research in tracking algorithms. In Fig. 4, we also observe that the baseline method is sensitive to the initialization and is not working robustly for a wide range of conditions. The poor generalization of many tracking algorithms has also been witnessed in visible-sensor tracking domain [12].



(a) *Frio10* (b) *Marathon-5*

Figure 4. Tracking results for single object tracking with tracking forward and tracking backward initializations

B. Tracking Multiple Objects from a Single View

The classic pipeline for tracking multiple objects from a single view involves object detection, temporal data association and object state estimation. These steps can be arranged in a sequential order, or placed in a batch processing mode. We refer to Wu's PhD thesis [13] for a detailed summary of the state-of-the-art algorithms. In addition to detecting noisy objects, the main challenge for thermal image analysis is to resolve the data association ambiguity in the presence of mutual occlusion. To address these problems, we collected five sequences: *Frio11*, and *Marathon-1, 2, 3, and 4* (Fig. 5). For the *Frio11* sequence, the task is to track all bats flying across the scene. For the *Marathon* sequences, the task is to track pedestrians, bicycles, motorcycles and cars. Two different viewpoints are provided with cropped images to focus on the region of interest. We annotated a single point for each bat in *Frio11* and a bounding box for each object in *Marathon*.

Baseline Method and Evaluation. For the bat sequence, the baseline method we adopted here is similar to the sequential tracking method proposed by Betke et al. [14]. This method detects bats by searching for the local maxima of each disconnected component after background subtraction. Then it sequentially associates detections to objects through bipartite matching and applies Bayesian filtering to estimate the motion of each bat. For the marathon sequences, we chose a batch processing method ("SDD-Net") [13] with a sparsity-driven object detector that handles mutual occlusions in the scene. The data association was implemented with the network flow formulation.

We use the popular "CLEAR MOT" metrics [15] to evaluate the tracking performance of our baseline methods. The Multiple Object Tracking Accuracy (MOTA) combines false positive detection rate, miss rate, and identity-switch rate into a single number with ideal value 100%; the Multiple Object Tracking Precision (MOTP) measures the average distance between ground truth and tracker output. For bounding box measurements, precision is defined according to the region overlap criterion with ideal value 1. For point measurements, it is based on the Euclidean distance with ideal value 0. To better assess the quality, we also report the numbers of Mostly Tracked (MT, $\geq 80\%$) trajectories, Mostly Lost (ML, $\leq 20\%$) trajectories, track fragmentations (FM, the number of times that a ground truth trajectory is interrupted), and identity switches (IDS, the number of times that a tracked trajectory changes its matched ground truth). These metrics depend on a user-defined threshold parameter that determines the hit/miss rates. A detection is a true positive if the distance between the detection and its matched ground truth is lower (or higher) than the threshold. We chose 0.5 for the region overlap threshold, and 15 pixels for the Euclidian distance threshold. The results of the baseline methods [14], [13] on two TIV test sequences are shown in Table III. It can be seen that the tracking algorithm achieves a low miss rate for the *Frio11* sequence because of the high contrast between foreground and background, but fails to handle frequent mutual occlusions and results in a high ID switch error. The noisy background makes the *marathon* sequences more challenging, as more than 10% objects were mostly lost. Clearly, even for state-of-the-art

Table III
RESULTS FOR MULTI-OBJECT TRACKING FROM A SINGLE VIEW.

Data	Method	MT	ML	FM	IDS	MOTA	MOTP
Frio11	Betke et al. [14]	96.9%	0.8%	410	1,222	65.0%	3.3 px
Marathon	SDD-Net [13]	60.9%	12.3%	172	158	62.1%	76.1%

algorithms, there is still large room to improve performance on these sequences.

C. Tracking Multiple Objects from Multiple Views

Occlusion, especially long-term occlusion, is one of the most difficult challenges in multi-object tracking. Occlusion reasoning may be improved if multiple cameras with overlapping fields of view can be provided. However, a new spatial data association step to establish the correspondence across the camera views must then be introduced [13]. We further classify the multi-object multi-view scenario into two categories: planar motion and free 3D motion. To address these two scenarios, we collected two sequences, *Atrium* and *Lab* for the planar motion; and *Velifer*, and *Davis08-sparse*, *Davis08-dense* and *Davis13-medium* for the free 3D motion. A few sample frames are shown in Fig. 6. We further provide camera calibration files that describe the multiview geometry of the scene, and annotations. We annotated a bounding box for the human sequences and a single point for each bat.

In the *Atrium* sequence, students are entering and leaving a building through an open area near the/exit doors. The students who are about to leave the building and those who just entered the building can be distinguished by their thermal profiles. The original video takes about 15 min but we removed all "idle" frames with no activity.

The *Lab* sequence captures interactions between 6 people walking close to each other. This sequence is more difficult to interpret because of a low camera view point and severe occlusion. Generally, with the help of the homography, it is easier to track objects on the ground plane than in the image plane. For the bat sequences, triangulation can be used to localize the animals in 3D space. The problem becomes more difficult as their density increases.

Baseline Method and Evaluation. For free 3D motion, we adopted the "Reconstruction-Tracking" algorithm [16], a variant of Multiple-Hypothesis-Tracking. This baseline method first reconstructs the 3D location of the object by solving a multi-dimensional assignment problem, and then sequentially tracks each object in 3D space. An enhanced version ("SDD-MHT"), proposed by Wu [13], applies a sparsity constrained optimization procedure to the reconstruction step above. The extension proves to be very effective to reduce the number of false positives, or "ghost points," which are generated by the false matches across camera views.

For the planar motion, we adopted the same sparsity driven object detector [13] to detect people on the ground plane

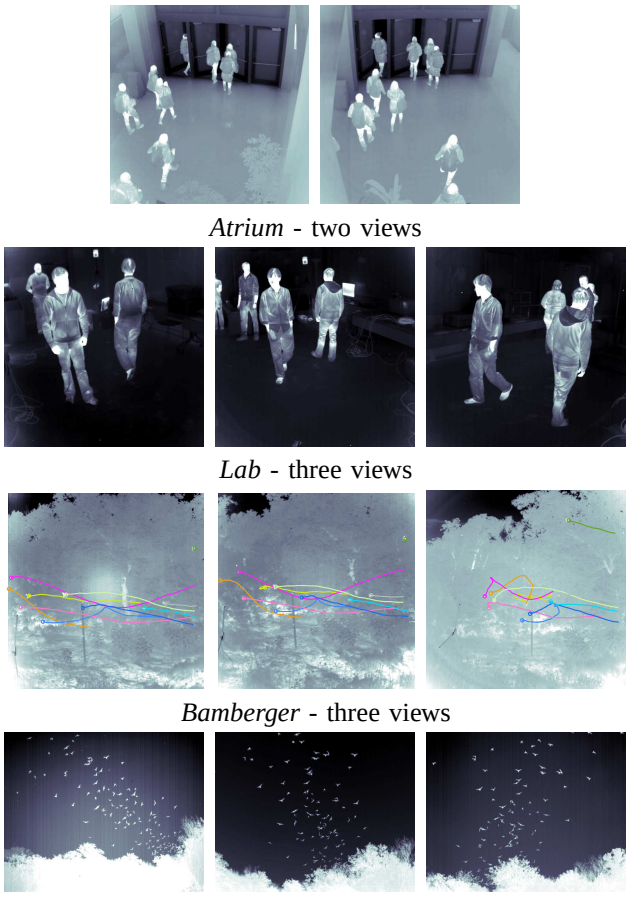


Figure 6. Sample frames for multi-object tracking from multiple views.

from the foreground estimation, as described in Section II-B. The data association step was based on Kalman filter and bipartite matching. Note that this baseline ("SDD-KF") and the variants of MHT above are all sequential online tracking methods.

To evaluate the tracking performance, the same "CLEAR MOT" metrics [15] were used. We chose 0.5 m on the ground plane as a miss/hit threshold for the localization of people, and 0.3 m in 3D space as the miss/hit threshold which approximates the physical size of a bat. Quantitative results for multi-object tracking from multiple views are listed in Table IV. As expected, the multi-camera setup helps when the 3D localization step can be solved accurately. Otherwise, more efforts are needed to improve the accuracy of 3D localization before the tracking step takes place,

Table IV
RESULTS FOR MULTI-OBJECT TRACKING FROM MULTIPLE VIEWS.

Data	Method	MT	ML	FM	IDS	MOTA	MOTP
Atrium-1-view	SDD-KF [13]	75.7%	2.4%	48	55	48.6%	72.5%
Davis08-sparse	MHT [16]	96.6%	0	105	97	64.1%	8.9 cm
Davis08-sparse	SDD-MHT [13]	95.2%	0	145	126	78.9%	5.7 cm
Davis08-dense	MHT [16]	71.9%	2.5%	274	355	-32.0%	10.0 cm
Davis08-dense	SDD-MHT [13]	61.1%	3.0%	454	444	44.9%	7.7 cm

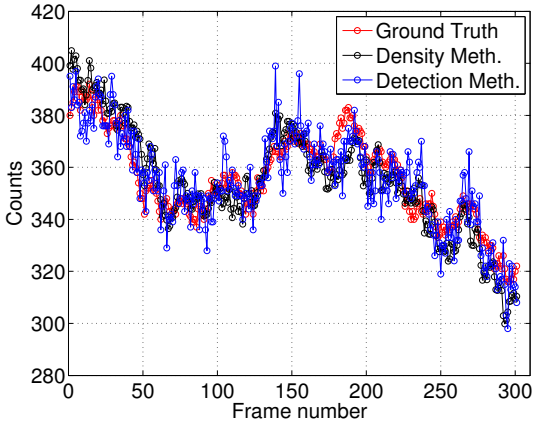
especially in dense tracking scenarios.

D. Visual Counting

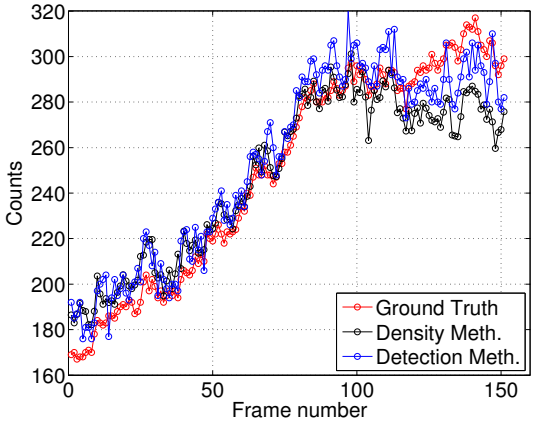
The visual counting task is to count the number of objects in the scene, a non-intrusive task for crowd analysis. One may think a straightforward solution is to apply object detection methods or even the multi-object tracking methods described in previous sections. However, it still remains a challenging problem to extend the scalability of these traditional techniques to handle very dense scenarios. Fortunately, techniques to count the objects without using an object detector exist [17], [18], [19]. To encourage research in this direction, we provide two sequences "Davis08-counting" and "Bracken-counting" to count the bats in a given region-of-interest, as shown in Fig. 7. For each frame, we only give the total number of bats as ground truth. We also provide a few training data that contain the location of every bat in an image.

Figure 7. Sample frames for counting.

Baseline Method and Evaluation. Counting methods can be broadly categorized by three classes: counting-by-detection, counting-by-regression, and counting-by-density-estimation. The counting-by-detection method typically needs a visual object detector to localize the object's position in the image. With the localization information, the counting is a trivial problem. The counting-by-regression method learns a regression model that directly maps some global features of the image to a number and for which it needs a large amount of training data. Finally, the counting-by-density-estimation method estimates the object density at each pixel based on local features, and integrates the density over the entire image. Here we report results on the two sequences with a customized bat detector [14] and a density estimation method [18]. The detector searches for the local maximum points, or key points, in each disconnected component after background subtraction. The final count is the total number of key points after non-maximum suppression. The density estimation method first computes dense SIFT features for the entire image and then approximates the density at each pixel by a linear transformation of the quantized SIFT feature. The final count is the integration of the density function over the image.



(a) Counting results on *Davis08-Counting*



(b) Counting results on *Bracken-Counting*

Figure 8. Comparison of frame-by-frame counting results of two baseline methods, the counting-by-density-estimation method (Density Meth.) and the counting-by-detection method (Detection Meth.) with the ground truth.

Frame-by-frame counting results on *Davis08-counting* and *Bracken-counting* are shown in Fig. 8. The mean numbers of objects per frame for these two sequences are 356 and 250, respectively. Both methods tended to underestimate

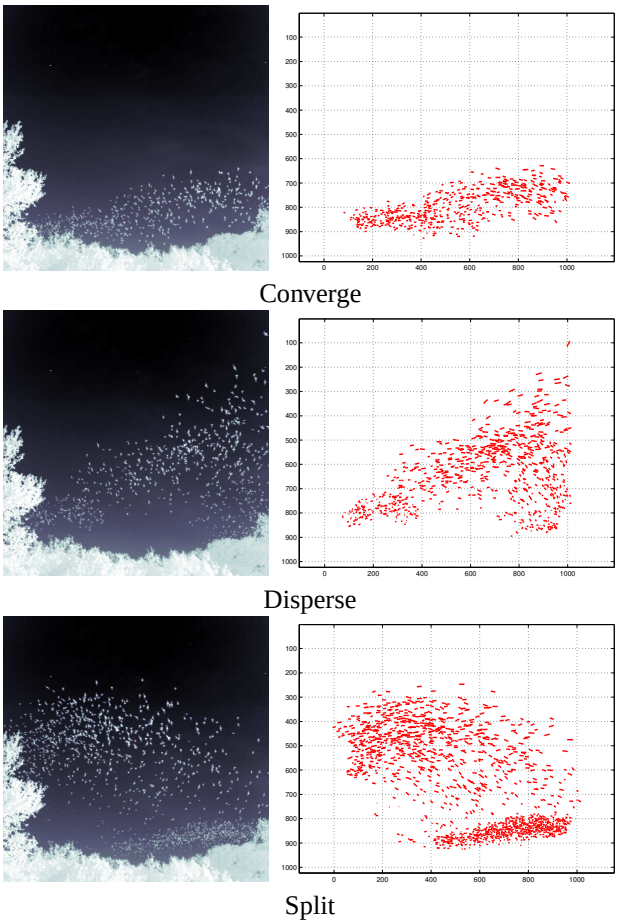


Figure 9. Sample frames for group motion estimation and flow vector annotations.

the number of objects in the crowds due to partial or even complete occlusion. No temporal information was used here. To evaluate the two methods, we computed the mean counting error over all frames as well as the standard deviation. The detection method achieved a 7.1 ± 5.8 error on *Davis08-counting* and a 10.4 ± 6.5 on *Bracken-counting*, while the density estimation method achieved 7.4 ± 5.0 and 11.8 ± 9.7, respectively. Given the fact that occlusion is difficult to resolve on image plane here, we believe it is promising to incorporate temporal information into the counting frameworks above.

E. Group Motion Estimation

Recent progress in visual event monitoring goes beyond the analysis in individuals. Crowd motion or group behavior studies have become popular in the computer vision community [20]. Unlike previous topics, one of the main challenges here is the actual lack of data as well as ground truth annotations. Meanwhile, researchers are still trying to devise interesting topics in different contexts and make formal problem definitions for them. Here we provide a long sequence, *Bracken-flow*, that shows part of the emergence of a Brazilian free-tailed bat colony. There are different group formations during the emergence, and we would like to continuously identify those motion patterns throughout the sequence using some group motion estimation method. Some unique patterns are shown in Fig. 9.

We manually divided the sequence *Bracken-flow* into multiple segments, each of which is associated with a motion pattern label. Some motion patterns repeat multiple times in the sequences. For each unique motion pattern, we also annotated the instantaneous flow vector (i.e., flow between 2 frames) for 10 examples. The annotations of the flow vector are noisy due to the high density of the objects in the scene. For a group behavior study, we are less interested in the accurate analysis of an individual. Instead, a high-level global description is desired. So the annotated flow vector here is only for reference purposes. This topic remains an open problem, and we expect to see algorithms dedicated to solving such problems in future.

III. Summary

With this paper, we introduced a new thermal infrared video dataset. We designed it to be extensive and diverse, and to include scenarios not present in existing datasets. Our intension is to provide a challenging benchmark for addressing several visual analysis tasks. We hereby publish the dataset, annotations, and the source code of the baseline methods with the evaluation protocols used to obtain the results in this paper and make them available to the public at `http://csr.bu.edu/BU-TIV/`.

Our preliminary study with this dataset showed that thermal infrared videos are not necessarily easier to process than data from visible sensors. We expect to see new ideas emerge in the future. Other researchers may design algorithms specifically for thermal infrared videos and improve the analysis results.

In near future, we plan to add data to the proposed benchmark. We will provide videos of additional scenes, as well as data from moving cameras, or camera network with non-overlapping fields of view.

Acknowledgment

This material is based upon work partially supported by Naval Research, grant N000141010952, and the National Science Foundation, grants 0910908 and 1337866. We would like to thank Brian Borucki, Mikhail Breslav, Qinxun Bai and other participants from Boston University for help with data collection.

References

[1] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, 2014.

[2] D. H. Theriault, N. W. Fuller, B. E. Jackson, E. Bluhm, D. Evangelista, Z. Wu, M. Betke, and T. L. Hedrick, "A protocol and calibration method for accurate multi-camera field videography," *The Journal of Experimental Biology*, February 2014.

[3] A. Torabi, G. Masse, and G.-A. Bilodeau, "An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 210–221, 2012.

[4] C. Conaire, N. O'Connor, E. Cooke, and A. Smeaton, "Comparison of fusion methods for thermo-visual surveillance tracking," in *Proc. IEEE Conf. Information Fusion*, 2006.

[5] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, "People detection and tracking from aerial thermal views," in *Proc. IEEE Conf. on Robotics and Automation*, 2014.

[6] J. Davis and M. Keck, "A two-stage approach to person detection in thermal imagery," in *Proc. IEEE Workshop on Applications of Computer Vision*, 2005.

[7] J. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 162–182, 2007.

[8] Z. Wu, A. Thangali, S. Sclaroff, , and M. Betke, "Coupling detection and data association for multiple object tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1948 – 1955.

[9] M. Schulz and L. Caldwell, "Nonuniformity correction and correctability of infrared focal plane arrays," *Infrared Physics and Technology*, vol. 36, no. 4, pp. 763–777, 1995.

[10] J. G. Harris and Y.-M. Chiang, "Nonuniformity correction of infrared image sequences using the constant-statistics constraint," *IEEE Trans. on Image Processing*, vol. 8, no. 8, pp. 1148–1151, 1999.

[11] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[12] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 1 – 8.

[13] Z. Wu, "Occlusion reasoning for multiple object visual tracking," Ph.D. dissertation, Boston University, USA, 2012.

[14] M. Betke, D. E. Hirsh, A. Bagchi, N. I. Hristov, N. C. Makris, and T. H. Kunz, "Tracking large variable numbers of objects in clutter," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[15] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, Feb. 2008.

[16] Z. Wu, N. I. Hristov, T. H. Kunz, and M. Betke, "Tracking-reconstruction or reconstruction-tracking? comparison of two multiple hypothesis tracking approaches to interpret 3d object motion from several camera views," in *Proc. IEEE Workshop on Motion and Video Computing (WMVC)*, Utah, December 2009.

[17] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Trans. Image Processing*, vol. 21, no. 4, pp. 2160–2177, 2012.

[18] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 1324–1332.

[19] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.

[20] B. E. Moore, S. Ali, R. Mehran, and M. Shah, "Visual crowd surveillance through a hydrodynamics lens," *Communication of ACM*, vol. 54, no. 12, pp. 64–73, 2011.